Milestone 2

Evan Edmunds

* Any surprises from your domain from these data?
  + The data for this project include 430 images of 9 types of fish in color. All the pictures are taken from the same angle, looking directly down at the profile of the fish against a blue background. The project will be using these images to train a classification model on each of the fish. Since this data seems to have only images of healthy adult fish, there are not many surprises to consider about the objective of the project.
* The dataset is what you thought it was?
  + As stated, the data includes 430 images of 9 types of fish in color. Each of these images are taken in the same context, so the goal will be to find what type of image processing technique will be able to classify the fish most effectively. All the images were stored in a central file directory, with one sub directory for each type of fish. All this data was imported into a pandas dataframe by first storing the pixel data in a numpy array and assigning the imaging data a tag based on what type of fish it was.
* Have you had to adjust your approach or research questions?
  + I have not had to change my approach or my research question. Since I assigned a tag to each instance of image data, this problem can be approached via a supervised learning algorithm. Since there has been no surprises as of yet, the central focus of the project has remained the same: construct an image recognition model that can be used for multiclass classification.
* Is your method working?
  + As of right now, the method is working. Doing some introductory analysis of the data, I found that the resolution for the images were split between 2832 x 2128 and 1024x768. To keep the size difference of these models from unduly influencing the model, I resized them both to 590 x 465. This maintained the same aspect ratio for both (4:3) as well as lowered the amount of data that would need to be fed into the algorithm later. Everything is now sitting in a Pandas dataframe, ready to undergo further analysis as well as preprocessing for the mode.
* What challenges are you having?
  + My challenge now is considering all the different possibilities for both preprocessing the images as well as the low number of images. Each fish only has 50 images, with one exception only having 30 images. To overcome this, I will have to augment the images so that I can have more data to work with for model training purposes. Due to generating more data, it may be necessary for me to look at the data in greyscale if the analysis and model training are prohibitively complex from a time standpoint. For image processing, due to the uniform background of all images, I think edge detection may be a useful approach, but I will explore other options as well. Efficiently generating more image data and finding the best way to process the data will be my next steps in continuing with the project.